Compiling a Corpus of Taiwanese Students' Spoken English

Lan-fen Huang
Language Centre, Shih Chien University
200 University Road
Nei-men, Kaohsiung 845 Taiwan
Lanfen.huang@gmail.com

Abstract

This paper reports the compilation of a corpus of Taiwanese students' spoken English, which is one of the twenty sub-corpora of the *Louvain International Database of Spoken English Interlanguage* (*LINDSEI*) (Gilquin et al., 2010). *LINDSEI* is one of the largest corpora of learner speech. The compilation process follows the design criteria of *LINDSEI* so as to ensure comparability across sub-corpora. The participants, procedures for data collection and process of transcription are all recorded. Sixty third- or fourth-year English majors in Taiwan are interviewed and recorded in English. Each interview is accompanied by a profile which contains information about such learner variables as age, gender, mother tongue, country, English learning context, knowledge of other foreign languages, amount of time spent in English-speaking countries and such interviewer variables as gender, mother tongue, knowledge of foreign languages and degree of familiarity with the interviewees. Another variable, the learners' English proficiency level based on the results of international standardised tests is collected; this is not available in other sub-corpora of *LINDSEI*. The participants' proficiency is similarly distributed across B1 to C1 levels in the Common European Framework of Reference. This paper concludes with a discussion of the contributions and research potential of the corpus.

1 Introduction

Corpus compilation, as it has developed, can be traced back to the 1960s (Sinclair, 1991). Research on corpora has mostly focused on written English and contributed a great deal of corpus-based grammatical description and explanation. In contrast, relatively few studies have emerged of corpora of spoken languages, which call for a time-consuming and laborious transcribing process. Yet it is widely acknowledged that this is an area which needs to be further explored (Carter and McCarthy, 1995). A similar trend is found in the field of learner corpora. Learner corpora have been used to study the written language of learners from different backgrounds, in terms of mother tongue. However, little research has been done on the spoken language produced by learners. One of the few major accomplishments in the corpus studies of learners' spoken English is the compilation of *the Louvain International Database of Spoken English Interlanguage* (*LINDSEI*) version 1 (Gilquin et al., 2010), which includes spoken English produced by learners from eleven different first languages (L1s). The present paper first introduces *LINDSEI* and then reports the compilation process of the Taiwanese sub-corpus, before discussing its contributions and potential for future research.

2 Overview of *LINDSEI*

The *LINDSEI* project began in 1995 and in 2010 published its first version, which includes sub-corpora of eleven L1s: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish and Swedish[1]. It involved 544 informal interviews and roughly one million tokens in total, with an average of 1,949 tokens in each one. About one third of the spoken data comes from the interviewers and two thirds from the learners.

In order to have comparable data across sub-corpora and to avoid the heterogeneity of interlanguage, the sub-corpora of *LINDSEI* must

[1] Another nine are in progress, including this Taiwanese sub-corpus. Please see *LINDSEI Partners* (Gilquin, 2012a) at http://www.uclouvain.be/en-307845.html (assessed on 22 August 2013).

27th Pacific Asia Conference on Language, Information, and Computation pages 199－205

meet an established set of criteria. Each corpus consists of 50 to 53 informal interviews between a learner and an interviewer. All learners are third- or fourth-year English-major students in countries where English is used as a foreign language and more than half the interviewers (64%) are native speakers (NSs) of English (Gilquin et al., 2010).

Each interview takes about 15 minutes to cover three tasks: set topics[2], free discussion and picture description. The first task serves as a warm-up activity. One of three topics is chosen by the interviewee. This lasts five to six minutes, including some follow-up questions put by the interviewer. The second task, taking seven to eight minutes, consists of free discussion of general topics, such as life at university, hobbies, travel experience, what the student hopes to do after university, family, etc. The object is not to stress and embarrass the interviewees with difficult questions but to get them to talk spontaneously. In the last few minutes, the interviewer asks the interviewee to look at a sequence of four pictures and tell the story that they illustrate. The student should not be given either the time or opportunity to make notes before describing the picture. It should be an improvised description.

All the interviews are orthographically transcribed and marked up according to the transcription guidelines (Gilquin, 2012b). Each transcription is accompanied by a profile which contains information about such learner variables as age, gender, mother tongue, country, English learning context, knowledge of other foreign languages, amount of time spent in English-speaking countries and such interviewer variables as gender, mother tongue, knowledge of foreign languages and degree of familiarity with the interviewees.

The eleven sub-corpora of *LINDSEI* offer a wide range of possibilities of research into Contrastive Interlanguage Analysis (CIA)[3]. The comparison can be done between different interlanguages as well as between any interlanguage and the native speech in the *Louvain Corpus of Native English Conversation* (*LOCNEC*), which is compiled by De Cock (2004), using the same structure as *LINDSEI*.

In addition, the written counterpart of *LINDSEI*, the *International Corpus of Learner English* (*ICLE*) (Granger et al., 2009) is a corpus of argumentative essays written by learners from sixteen L1 backgrounds. *LINDSEI* and *ICLE* share ten mother tongue backgrounds, which makes it possible to compare spoken and written interlanguages.

3 Taiwanese Sub-corpus of Spoken English

The compilation of the Taiwanese sub-corpus of *LINDSEI* began in October 2012 and went on for one year, sponsored by the National Science Council, Taiwan, under grant number NSC101-2410-H-158-012.

3.1 Recruitment of Participants

The participants were 60[4] third- or fourth-year undergraduate students majoring in English in the six universities in Taiwan, listed in Table 1 below.

	University	Number of participants
1	Shih Chien University	7
2	Wenzao Ursuline College of Languages	10
3	National Cheng Kung University	16
4	National Pingtung University of Education	12
5	National Taiwan University of Science and Technology	9
6	National Kaohsiung University of Applied Sciences	6
	Total	60

Table 1. Universities participating in the Taiwanese sub-corpus of *LINDSEI*

[2] The three set topics are: 1) *An experience you have had which has taught you an important lesson. You should describe the experience and say what you have learnt from it.* 2) *A country you have visited which has impressed you. Describe your visit and say why you found the country particularly impressive.* 3) *A film/play you've seen which you thought was particularly good/bad. Describe the film/play and say why you thought it was good/bad* (Gilquin et al., 2010, p. 8).

[3] The term, Contrastive Interlanguage Analysis (CIA) was coined by Granger (1996; 1998).

[4] The *LINDSEI* team requires all contributors to a sub-corpus to submit 50 recordings and their accompanying profiles. In case of problems such as unintelligible sound quality or an incomplete learner profile for any of the contributors, 60 recordings were made. 50 out of the 60 learners will be sent to the *LINDSEI* team, who will further process them. Therefore, the data in the Taiwanese sub-corpus of *LINDSEI* reported in this paper will differ slightly from the final version included in the second version of *LINDSEI*.

The participants were recruited through an advertisement on campus or at the invitation of their instructors. They were informed that the collected spoken data would be used for research purposes and had to give their permission by signing a learner profile questionnaire (see Appendix A) on the day of the interview. The questionnaire used for the Taiwanese corpus was slightly adapted from that in *LINDSEI* by adding one question: *Have you ever taken an English proficiency test? If yes, please give the name of the test, your result and date of the test*. Most of the learners gave their TOEIC scores, but some had IELTS, TOEFL, BULATS, GEPT and CSEPT grades [5]. Table 2 below lists the distribution of the 60 learners' English proficiency in the four levels of the Common European Framework of Reference (CEFR). The learners' proficiency is similarly distributed across the B1 to C1 levels; therefore, it is best described as ranging from intermediate to advanced. The Taiwanese sub-corpus is similar to other sub-corpora in *LINDSEI*. Although information about the learners' proficiency in *LINDSEI* was not available, a tentative study, based on a random sample of five learners from each sub-corpus, indicates that 64% were rated as high-intermediate (and lower) and 36% advanced (Gilquin et al., 2010, pp. 10-11).

Level	Number	Percentage
B1	14	23.3%
B2	18	30.0%
C1	19	31.7%
C2	1	1.7%
n/a	8	13.3%
Total	60	100%

Table 2. The distribution of the 60 learners' English proficiency in the four levels of CEFR

Four interviewers, one American, one British and two Taiwanese teachers of English, were involved in the data collection (see Table 3). Ideally, the interviewers should have been NSs of English, since it may be easier to develop natural communication when the learners talk with someone who does not share the same L1. However, to fit in with the availability of the interviewers who were NSs, the learners and the researcher, 70% of interviews were done by NSs and the remainder by Taiwanese teachers of English. They were briefed beforehand on how to conduct the interview and fully aware of the use of the transcripts and audio files for research purposes.

Inter-viewer	Gender	Mother tongue	Number of interviews (Percentage)	Transcript Number
1	Male	British English	22 (36.7%)	TW011-032
2	Male	American English	20 (33.3%)	TW001-010 TW033-042
3	Male	Chinese	9 (15.0%)	TW043-051
4	Female	Chinese	9 (15.0%)	TW052-060
			60 (100%)	

Table 3. The interviewers' gender and mother tongue

3.2 Procedures for Informal Interviews

On the day of the interview, the learners of English were asked to fill in a profile questionnaire (Appendix A), with the assistance of the researcher. This form included information about learner variables and was signed and dated to signify written consent to use the recorded interviews for research purposes. In order to make the best use of time without keeping the interviewers waiting, this task of filling the questionnaire was done by some learners after the interviews. Either way, the learners were well aware of being recorded.

After filling in the questionnaires, the learners were given at least five minutes to prepare to talk on one of the three set topics. Then, the learners were invited to enter a classroom or meeting room where two electronic recorders had been set up. The researcher left the room as soon as she had made sure that the recorders were working, because the students might have felt under pressure if two people had been listening to them.

As reported in the previous section, the whole informal interview took about 15 minutes. During this period, the interviewer tried his/her best to be friendly and to help students talk more by giving quick responses and specific questions, and the learners were given neither the time nor

[5] The Test of English for International Communication (TOEIC), International English Language Testing System (IELTS), Test of English as a Foreign Language (TOEFL), and Business Language Testing Service (BULATS) are internationally recognised certificates. The General English Proficiency Test (GEPT) and College Student English Proficiency Test (CSEPT) are locally developed tests in Taiwan.

the opportunity to write notes. This interview aimed to collect spontaneous speech from the learners.

After the interviews, the learners were given a voucher for NT$200 (US$1 equals NT$30) to spend. The recordings and learner profiles were coded for the transcribing process.

3.3 Process of Transcription

The 60 interviews were orthographically transcribed and marked up by two research assistants following the guidelines provided by the *LINDSEI* project (Gilquin, 2012b). The transcription work for a 15-minute interview might take five to ten hours, depending on the transcribers' experience of transcribing. The two transcribers spent more time to begin with, when they were not yet very familiar with the transcription guidelines. All the transcripts were double-checked by the researcher. Each of them took about 30 to 60 minutes to finish.

In the process of transcription, two pieces of computer software were used, *Audacity* (2013) and *Windows Media Player*. *Audacity* was used to edit the sound recordings, in particular for deleting redundant time at the beginning and end of the interviews. It also made it possible to manipulate the sound file, e.g. by reducing its speed, playing it back several times, etc.

The task of orthographic transcribing needed less skill. The mark-up process required more training. Of the twenty aspects of transcription in the guidelines, the marking-up of overlapping speech was most difficult and time-consuming.

4 Contributions of the Taiwanese sub-corpus of *LINDSEI*

The establishment of the Taiwanese learner corpus of spoken English will make contributions in three ways: 1) by increasing the visibility of Taiwanese learners in the international academia; 2) by informing the teaching of spoken English to Taiwanese students; and 3) by serving as a model for the compilation of corpora of spoken English in Taiwan.

First, Taiwanese learners represent one group of Chinese speakers, as well as the Chinese sub-corpus compiled in mainland China, in the fields of corpus studies and interlanguage research. *LINDSEI* is currently the most comprehensive learner corpus project and includes international collaboration from twenty groups. Being one of the sub-corpora of *LINDSEI*, without doubt, increases the visibility of Taiwan in international academia and contributes to the research on spoken English. The spoken data collected in Taiwan will be shared with other groups of L1s. This, compared with a self-designed learner corpus, enables researchers worldwide to conduct a wider range of investigations. Furthermore, the learner speech collected in Taiwan in 2012 and 2013 offers the most recent data of this kind, while those in the Chinese sub-corpus were compiled in 2001 (Gilquin et al., 2010). The information in the learner profiles of the Chinese sub-corpus shows that 48 out of 53 learners (90.6%) had received six years of English education at school before they began their first degree and none of the learners had ever stayed in an English-speaking country. By contrast, the learners in the Taiwanese sub-corpus had much greater exposure to English. They had on average nearly ten years of English learning before entering university and 21 out of 60 (35%) learners had stayed in countries where English is spoken for an average of 6.8 months.

Second, the usage patterns of Taiwanese learners can be identified to facilitate and improve the teaching of spoken English. The importance of corpus studies and applications has been stated in recent international conferences on Applied Linguistics held in Taiwan (e.g. the 18th International Symposium on English Teaching: Internet- and Corpus-based English Instruction (13-15 November 2009), the 2012 International Conference on Applied Linguistics and Language Teaching: Technological and Traditional Teaching and Learning (19-21 April 2012), and the 2012 LTTC International Conference: The Making of a Translator (28-29 April 2012)). However, there has hitherto been no learner corpus of spoken English available for research purposes. It is worth noting that the Language Training and Testing Centre in Taiwan has undertaken to transcribe the speaking tests of GEPT, which was developed in Taiwan, but it might take some time for the learner corpus to be published. In mainland China, some learner corpora have been made available, for example, the *Spoken and Written English Corpus of Chinese Learners*, version 1.0 (Wen et al., 2005) and version 2.0 (Wen et al., 2008); and the *Chinese Learner Spoken English Corpus* (Yang and Wei, 2005). The data in these corpora were collected from speaking tests which involve retelling a story, describing a picture and discussing a topic. In the test-taking context, learners' speech was

restricted and unnatural. In contrast, the spoken English produced in the informal interviews for *LINDSEI* was relatively authentic. The learners were voluntary and the setting was outside the classroom and not exam-oriented.

Third, this corpus will be the first publicly available learner corpus in Taiwan. It will serve as a model for the compilation of corpora. In Taiwan, the development of corpus studies is still in its infancy. This project, in collaboration with the *LINDSEI* team in Belgium, provides research training for the researcher as well as the team members. The researcher benefits from interacting with international researchers in the field of Corpus Linguistics and from being involved in the process of transcribing, which is seen as an analytical tool (Swann, 2010). Both these advantages will help the researcher to exploit the potential of the collected data. The team members gain research experience and broaden their scope in the expectation that more corpus studies will be done in future.

5 Research Possibilities

The corpus of Taiwanese students' spoken English provides a range of possibilities for research. As mentioned in Section Two, the sub-corpora in *LINDSEI* have been employed in CIA, in which two types of comparison can be made: 1) between NS and learner languages (in this case, *LOCNEC* (De Cock, 2004) and the Taiwanese sub-corpus) and 2) between speakers of different mother tongues (the Taiwanese sub-corpus and any other sub-corpora of *LINDSEI*). There is a growing interest in quasi-longitudinal studies, i.e. comparing learners of the same L1 at different proficiency levels. Information about learners' English proficiency levels is available (see Table 2) and reliable, because it based on the results of international standardised tests of English proficiency. In both CIA and quasi-longitudinal studies, a number of investigations can be pursued, such as lexis, phraseology, organization of spoken discourse, and features of spoken English.

Among the five features of spoken English 1) deictic expressions, 2) situational ellipsis, 3) headers, tails and tags, 4) discourse markers and 5) polite and indirect language, vague language and approximations (Carter and McCarthy, 2006), discourse markers have attracted much research attention (e.g. on Chinese learners: He and Xu, 2003; Fung and Carter, 2007; Liu, 2010; Huang, 2011). The quantitative corpus studies have revealed the usage of discourse markers by learners. Such research has been conducted across the eleven sub-corpora by Gilquin and Granger (2011; forthcoming). These researchers point out that using *LINDSEI* as an aggregate may conceal variations between learners of different L1s as well as between learners in one specific corpus. It seems that the L1 plays an important role for ESL learners.

In terms of practical applications, the learner corpus research has certainly helped us to improve our understanding of learner language and to inform English Language Teaching. However, there is always more work to do. As De Cock (2010) notes in her call for more studies using spoken learner corpora in the classroom, the compilation of the Taiwanese sub-corpus of *LINDSEI* will certainly facilitate research on Chinese-speaking learners, which is one of the biggest groups to use English as a foreign language.

Acknowledgments

This work was supported by the National Science Council, Taiwan, under grant number NSC101-2410-H-158-012. Without this funding, the Taiwanese sub-corpus of *LINDSEI* would not have been possible. My gratitude goes to the *LINDSEI* team at the Centre for English Corpus Linguistics of the Université Catholique de Louvain, Belgium, in particular, the project leader, Prof Sylviane Granger and the coordinator, Dr Gaëtanelle Gilquin. The efforts of my project team members, Ms Hsiao-hui Lin, Ms Miranda Yu-ting Huang, Dr Jon Nichols, Mr Simon Kubelec, Mr Alex Jou and Mr Chih-hao Hsueh are most appreciated. Special thanks are due to my contacts in the six universities participating in this corpus and the Taiwanese learners who agreed to be interviewed and recorded.

References

2013 members of the Audacity development team 2013. Audacity (Version 2.0.3). Available at http://audacity.sourceforge.net/.

Anping He and Manfei Xu. 2003. Small words in Chinese EFL learners' spoken English. Foreign Language Teaching and Research, 35(6), 446-453.

Binmei Liu. 2010. Discourse Marker Use by L1 Chinese EFL Speakers. (PhD thesis), University of Florida.

Gaëtanelle Gilquin and Sylviane Granger. 2011. The use of discourse markers in corpora of native and learner speech: From aggregate to individual data. Paper presented at the Corpus Linguistics Conference 2011, Birmingham.

Gaëtanelle Gilquin, Sylvie De Cook, and Sylviane Granger (Eds.). 2010. LINDSEI Louvain International Database of Spoken English Interchange. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Gaëtanelle Gilquin. 2012a. LINDSEI Partners. Retrieved 26 August, 2013, from http://www.uclouvain.be/en-307845.html

Gaëtanelle Gilquin. 2012b. Transcription guidelines. Retrieved 26 August, 2013, from http://www.uclouvain.be/en-307849.html

Gaëtanelle Gilquin, and Sylviane Granger. forthcoming. Learner language. In D. Biber and R. Reppen (Eds.), *Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press.

Joan Swann. 2010. Transcribing spoken interaction. In S. Hunston and D. Oakey (Eds.), Introducing applied linguistics: Concepts and skills (pp. 163-176). Abingdon: Routledge.

John Sinclair. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Lan-fen Huang. 2011. Discourse markers in spoken English: A corpus study of native speakers and Chinese non-native speakers. (PhD thesis), University of Birmingham, UK. Retrieved from http://etheses.bham.ac.uk/2969/

Loretta Fung and Ronald Carter. 2007. Discourse markers and spoken English: Native and learner use in pedagogic settings. Applied Linguistics, 28(3), 410-439.

Ronald Carter and Michael McCarthy. 1995. Grammar and the spoken language. Applied Linguistics, 16(2):141-158.

Ronald Carter and Michael McCarthy. 2006. Cambridge Grammar of English. Cambridge: Cambridge University Press.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot (Eds.). 2009. International Corpus of Learner English (2nd ed.). Louvain-la-Neuve: Presses universitaires de Louvain.

Sylviane Granger. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg and M. Johansson (Eds.), Languages in Contrast. Text-based cross-linguistic studies. Lund Studies in English 88 (pp. 37-51). Lund: Lund University Press.

Sylviane Granger. 1998. The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), Learner English on Computer (pp. 3-18). Harlow: Longman.

Sylvie De Cock. 2004. Preferred sequences of words in NS and NNS speech. Belgian Journal of English Language and Literatures, New Series 2:225-246.

Sylvie De Cock. 2010. Spoken learner corpora and EFL teaching. In M. C. Campoy, B. Bellés-Fortuño and M. L. Gea-Valor (Eds.), Corpus-based Approaches to English Language Teaching (pp. 123-137). London: Continuum.

Appendix A. Learner Profile (adapted from Gilquin et al., 2010, pp. 110-111)

LEARNER PROFILE

==========

Text code: (to be filled in by the researcher)

==========

Surname: **First name(s)**:

Age:

Male ☐ Female ☐

Nationality:

Country:

Native language:

Father's mother tongue:

Mother's mother tongue:

Language(s) spoken at home: (if more than one, please give the average % use of each)

Education:

Primary school - medium of instruction:

Secondary school - medium of instruction:

Current studies:

Current year of study:

Institution:

Medium of instruction:

English only ☐

Other language(s) (specify) _____ ☐

Both ☐

==========

Years of English at school:

Years of English at university:

Stay in an English-speaking country:

Where?

When?

How long?

Have you ever taken an English proficiency test? If yes:

Name of the test:

Result: Date:

==========

Other foreign languages in decreasing order of proficiency:

==========

I hereby give permission for my interview to be used for research purposes.

Date: ..... Signature: .....

****

Section to be filled in by the interviewer

Interviewer: Male ☐ Female ☐

Native language:

Foreign languages (in decreasing order of proficiency):

Relation with learner: Familiar ☐ Vaguely familiar ☐ Unfamiliar ☐

(If possible, please be more specific, e.g. learner's professor, TA, etc: ……..……..)